Assessing Chronic Stress, Coping Skills, and Mood Disorders through Speech Analysis: A Self-Assessment 'Voice App' for Laptops, Tablets, and Smartphones

Braun, Silke ; Annovazzi, Chiara ; Botella, Cristina ; Bridler, René ; Camussi, Elisabetta ; Delfino, Juan P ; Mohr, Christine ; Moragrega, Ines ; Papagno, Costanza ; Pisoni, Alberto ; Soler, Carla ; Seifritz, Erich ; Stassen, Hans H

Abstract: Background. Computerized speech analysis (CSA) is a powerful method that allows one to assess stress-induced mood disturbances and affective disorders through repeated measurements of speaking behavior and voice sound characteristics. Over the past decades CSA has been successfully used in the clinical context to monitor the transition from "affectively disturbed" to "normal" among psychiatric patients under treatment. This project, by contrast, aimed to extend the CSA method in such a way that the transition from "normal" to "affected" can be detected among subjects of the general population through 10-20 self-assessments. Methods. Central to the project was a normative speech study of 5 major languages (English, French, German, Italian, Spanish). Each language comprised 120 subjects stratified according to gender, age, and education with repeated assessments at 14-day intervals (total n=697). In a first step, we developed a multivariate model to assess affective state and stress-induced bodily reactions through speaking behavior and voice sound characteristics. Secondly, we determined language-, gender-, and age-specific thresholds that draw a line between "natural fluctuations" and "significant changes". Thirdly, we implemented the model along with the underlying methods and normative data in a self-assessment "voice app" for laptops, tablets, and smartphones. Finally, a longitudinal self-assessment study of 36 subjects was carried out over 14 days to test the performance of the CSA method in home environments. Results. The data showed that speaking behavior and voice sound characteristics can be quantified in a reproducible and language-independent way. Gender and age explained 15-35% of the observed variance, whereas the educational level had a relatively small effect in the range of 1-3%. The self-assessment "voice app" was realized in modular form so that additional languages can simply be "plugged-in", once the respective normative data become available. Results of the longitudinal self-assessment study in home environments demonstrated that CSA methods work well under most circumstances. Conclusions. We have successfully developed and tested a self-assessment CSA method that can monitor transitions from "normal" to "affected" in subjects of the general population in the broader context of mood disorders. Our easy-to-use "voice app" evaluates sequences of 10-20 repeated assessments and watches for affect- and stress-induced deviations from baseline that exceed language-, gender-, and age-specific thresholds. Specifically, the "voice app" provides users with stress-related "biofeedback" and can help to identify that 10-15% subgroup of the general population that exhibits insufficient coping skills under chronic stress and may benefit from early detection and intervention prior to developing clinically relevant symptoms.

DOI: https://doi.org/10.1159/000450959

Original Paper

Psychopathology

Psychopathology 2016;49:406–419
DOI: 10.1159/000450959

Received: March 24, 2016
Accepted after revision: September 19, 2016
Published online: November 15, 2016

Assessing Chronic Stress, Coping Skills, and Mood Disorders through Speech Analysis: A Self-Assessment 'Voice App' for Laptops, Tablets, and Smartphones

Silke Braun[a] Chiara Annovazzi[d] Cristina Botella[e] René Bridler[a] Elisabetta Camussi[d] Juan P. Delfino[a] Christine Mohr[c] Ines Moragrega[e] Costanza Papagno[d] Alberto Pisoni[d] Carla Soler[e] Erich Seifritz[b] Hans H. Stassen[a]

[a]Institute for Response-Genetics, Psychiatric University Hospital (KPPP), University of Zurich, and [b]Psychiatric University Hospital (KPPP), Zurich, and [c]Institute of Psychology, University of Lausanne, Lausanne, Switzerland; [d]Dipartimento di Psicologia, Università di Milano-Bicocca, Milan, Italy; [e]Clinical Psychology, University of Jaume I, Castellon, Spain

Key Words

Chronic stress · Coping behavior · Affective reactions · Voice sound characteristics · Speaking behavior · Self-assessment · Depressive symptoms · Early detection · Prevention · Biofeedback

Abstract

Background: Computerized speech analysis (CSA) is a powerful method that allows one to assess stress-induced mood disturbances and affective disorders through repeated measurements of speaking behavior and voice sound characteristics. Over the past decades CSA has been successfully used in the clinical context to monitor the transition from 'affectively disturbed' to 'normal' among psychiatric patients under treatment. This project, by contrast, aimed to extend the CSA method in such a way that the transition from 'normal' to 'affected' can be detected among subjects of the general population through 10–20 self-assessments. ***Methods:*** Central to the project was a normative speech study of 5 major languages (English, French, German, Italian, and Spanish). Each language comprised 120 subjects stratified according to gender, age, and education with repeated assessments at 14-day intervals (total n = 697). In a first step, we developed a multivariate model to assess affective state and stress-induced bodily reactions through speaking behavior and voice sound characteristics. Secondly, we determined language-, gender-, and age-specific thresholds that draw a line between 'natural fluctuations' and 'significant changes'. Thirdly, we implemented the model along with the underlying methods and normative data in a self-assessment 'voice app' for laptops, tablets, and smartphones. Finally, a longitudinal self-assessment study of 36 subjects was carried out over 14 days to test the performance of the CSA method in home environments. ***Results:*** The data showed that speaking behavior and voice sound characteristics can be quantified in a reproducible and language-independent way. Gender and age explained 15–35% of the observed variance, whereas the educational level had a relatively small effect in the range of 1–3%. The self-assessment 'voice app' was realized in modular form so that additional languages can simply be 'plugged in' once the respective normative data be-

KARGER

E-Mail karger@karger.com
www.karger.com/psp

0254–4962/16/0496–0406$39.50/0

Hans H. Stassen
Institute for Response-Genetics
Psychiatric University Hospital (KPPP)
CH–8032 Zurich (Switzerland)
E-Mail k454910@bli.uzh.ch

come available. Results of the longitudinal self-assessment study in home environments demonstrated that CSA methods work well under most circumstances. ***Conclusions:*** We have successfully developed and tested a self-assessment CSA method that can monitor transitions from 'normal' to 'affected' in subjects of the general population in the broader context of mood disorders. Our easy-to-use 'voice app' evaluates sequences of 10–20 repeated assessments and watches for affect- and stress-induced deviations from baseline that exceed language-, gender-, and age-specific thresholds. Specifically, the 'voice app' provides users with stress-related 'biofeedback' and can help to identify that 10–15% subgroup of the general population that exhibits insufficient coping skills under chronic stress and may benefit from early detection and intervention prior to developing clinically relevant symptoms.

Introduction

Chronic stress can lead to serious health problems and can affect nearly every system of the human body, as suggested by physical, cognitive, affective, and behavioral symptoms. Indeed, for a certain percentage of the general population, chronic stress raises blood pressure, increases the risk of heart attack and stroke, suppresses the immune system, and increases the vulnerability to psychiatric disorders [1, 2]. Health surveys indicate that the stress-induced burden is closely related to a pronounced lack of coping skills which obviously can let things escalate in the long run [see for example 3]. Of particular interest, therefore, are tools that help to identify that 10–15% subgroup of the general population that exhibits insufficient coping skills under chronic stress and may benefit from early detection and intervention prior to developing clinically relevant symptoms.

Human speech is greatly influenced by the speaker's affective state, which reflects feelings like chronic stress, sadness, happiness, grief, guilt, fear, anger, aggression, faintheartedness, shame, love, or doziness – and, occasionally, by depressive or psychotic symptoms [4–6]. Indeed, attentive listeners discover a lot about the mood and affective state of their dialog partners without having to talk about it explicitly during a conversation or on the phone. For example, hectic and abrupt, or delayed and monotonous speech may indicate stress-related or affective problems, provided such behavior persists over a longer time.

Our interest in speech analysis has a psychiatric background because speech dysfunctions, such as slow, delayed or monotonous speech, are prominent features of severe mental disorders, for example, major depressive disorders, schizoaffective disorders, or schizophrenic disorders. Clinicians routinely monitor speaking behavior and voice sound characteristics among affectively disturbed patients as indicators of clinical change. Given the clinical relevance of changes in speaking behavior and voice sound characteristics for the patients' time course of recovery, we have developed a computerized speech analysis (CSA) method that allows one to monitor the transition from 'affected' to 'normal' among patients through a series of repeated assessments over 2–3 weeks [7, 8]. For the development of the CSA method, patients suffering from major depressive disorders were assessed at 2-day intervals by means of standardized psychopathology ratings (observer ratings) and 2- to 3-min speech recordings, with raters being 'blind' to speech-recording results and speech-recording staff 'blind' to psychopathology rating results.

Among the patients suffering from major depressive disorders we found essentially parallel time courses of recovery for the Hamilton Depression Scores (HAMD-17) on the one hand and voice sound characteristics on the other. The sample was comprised of 'early responders' with an onset of improvement within the first 10 days of treatment, 'late responders' with an onset of improvement after day 10 of treatment, and 'non-responders' who did not respond to treatment[1]. In 65–75% of cases HAMD-17 scores and voice sound characteristics showed a close correlation of $r \geq 0.8$ in single-case analyses.

Clearly, this method can also be used to monitor the transition from 'normal' to 'affected' among subjects at risk of affective disorders. The project 'Early Detection and Prevention of Mood Disorders' aimed to translate our patient-oriented, lab-based CSA method into a self-assessment 'voice app' for laptops, tablets, and smartphones such that the resulting 'voice app' (1) does not need special equipment, (2) covers several major languages, (3) has been trained to deal with the various forms of missing data, (4) blanks out acoustical artifacts that occasionally show up in self-assessment recordings, (5) analyzes 10–20 repeated assessments (up to 31) in order to monitor speaking behavior and voice sound characteristics as a function of time, (6) detects deviations from 'normality', and (7) can readily be used by

[1] The specifics pertinent to the time course of recovery from major depressive disorders are detailed in Stassen et al. [8].

everyone for highly informative self-assessments at very low costs.

Primary target populations of the project are (1) patients recovered from major depressive disorders and at risk of relapse, and (2) college and university students in the age range of 17–24 years. In fact, college and university students encounter significant levels of chronic stress over quite a long time, which can aggravate preexisting psychiatric conditions or can trigger the development of new mental health problems or other noncommunicable diseases. Academic and nonacademic stresses include competition in classroom, tight schedules, frequent exams, moving away from home, adaptation to new social environment, financial issues, and transition to a new developmental stage – nota bene in a period of life where 75% of subjects with major psychiatric disorders have their onset (age range of 17–24 years) [9]. Specifically, insufficient coping behavior under chronic stress can significantly affect the academic performance of students, can lead to elevated alcohol consumption and/or the use of illegal drugs, and can cause premature withdrawal from college or university prior to the completion of education.

Our CSA method follows a single-case approach with a series of repeated self-assessments at prespecified intervals. Each individual subject serves as his/her own reference, and the method looks for relative changes that exceed language-, gender-, and age-specific thresholds derived from representative normative samples [10]. Though short-term, often stress-induced fluctuations in mood are constituents of human life, significant deviations from baseline that persist over a longer time period (>2 days) might make it necessary to do something about it.

In this context it is important to note that we do *not* use the self-assessment CSA method to classify psychological states or psychiatric disorders nor to derive clinical diagnoses. In fact, attempts to derive a clinical diagnosis from single voice recordings – as sometimes proposed in the literature – can be grossly misleading. Even though the standard case-control approach with 'depressive' patients readily yields 75–85% of correctly classified subjects for almost any speech parameter set and classification method, a closer look reveals that more than 40% of the achievable discriminating power is related to the patients' medication and side effects. Also, there is a substantial overlap in terms of speaking behavior and voice sound characteristics between patients with major depressive disorders, schizoaffective disorders, and schizophrenic disorders. In consequence, false-positive and false-negative classification errors of such approaches typically exceed 20%.

By contrast, the interactive 'voice app' analyzes 10–20 repeated assessments (up to 31, preferably at 1-day intervals) in order to monitor speaking behavior and voice sound characteristics as a function of time. Results are presented in such a way that they are directly interpretable by almost every user and can be used to better cope with the ups and downs of the daily grind ('biofeedback': users can learn, for example, to control stress-related bodily reactions). Getting involved and doing something about it is the most important step for people under elevated risk of mood disorders. And this is what prevention is all about.

Methods

Normative Study: 697 Test Persons and 5 Major Languages

Central to the project was a normative study comprised of 5 substudies following the same experimental design and carried out with healthy volunteers in Bristol (English: n = 117), Lausanne (French: n = 128), Zurich (German: n = 208), Milan (Italian: n = 120), and Valencia (Spanish: n = 124). The chosen study sites with 2 'stress-timed' languages (English, German) and 3 'syllable-timed' languages (French, Italian, Spanish) allowed for comparisons within and between 'stress-timed' and 'syllable-timed' languages [for details on prosody, see 11, 12].

Samples were stratified according to gender, age (4 age groups: 18–30, 31–40, 41–50, and 51–65 years), and education (4 categories: remedial, junior high, high, and college or university). The test persons were asked to fill out the 63-item Zurich Health Questionnaire (ZHQ)[2] which assesses 'regular exercises', 'consumption behavior', 'impaired physical health', 'psychosomatic disturbances', and 'impaired mental health' (http://www.ifrg.uzh.ch/instruments.php). Thus, subjects with a previous history of psychiatric disorders could be excluded.

Speech signals were recorded as time series with a sampling rate of 96 kHz at a 16-bit resolution. The test persons were invited to present 4 types of speech, twice at 14-day intervals, at a fixed time

[2] The ZHQ was developed by Kuny and Stassen [24] and was used, for example, as external validation in our coping behavior studies with 2,520 students from 6 colleges and universities in the US, South America, and Europe [1, 2]. The results of these studies consistently showed highly significant correlations ($p < 0.0001$) across study sites between insufficient coping behavior under chronic stress on the one hand and the ZHQ factors 'impaired physical health', 'psychosomatic disturbances', and 'impaired mental health' on the other. By contrast, insufficient coping behavior under chronic stress and 'regular exercises' were found to be inversely related to each other. Moreover, we consistently found ZHQ alcohol consumption and ZHQ regular exercises to have a highly significant influence on voice sound ($p < 0.0003$). Regarding general health, speaking behavior and voice sound characteristics explained, for example, 24.9% of the observed variance in the test persons' state of ZHQ mental health. Regarding biological quantities, the ZHQ parameters age, height, and weight were found to explain between 39.7 and 45.4% of the observed variance in voice sound.

in the morning, according to the following scheme: (1) counting out loud from 1 to 40, (2) reading out loud an emotionally neutral passage of a children's book, (3) reading out loud an emotionally stimulating passage from a famous novel, and (4) counting out loud again from 1 to 40. The entire recording procedure took 10 min including individual volume calibration. The recordings were carried out in acoustically shielded rooms, using high-end microphones (Sennheiser MKH40 P48), along with A/D converters featuring 0.1dB linearity.

Additionally, we carried out a longitudinal self-assessment study in order to test the performance of the CSA method in home environments. Test persons were recruited in Zurich from students under exam stress (n = 18; German: 'stress-timed' language), and in Valencia from unemployed adults (n = 18; Spanish: 'syllable-timed' language). They received a low-cost netbook in combination with a low-cost microphone, were instructed on how to use the system, and were asked to perform speech recordings every afternoon in their home environments over a period of 14 days.

Parameter Extraction

Speech production is the result of a joint effort of mind and body. It involves a cascade of steps from utterance planning to final sound production with hundreds of degrees of freedom. Rhythm, stress, and intonation ('prosody') greatly influence the verbal and nonverbal content of the transmitted speech. Despite this complexity, speech characteristics can be roughly described by a few major features: speaking behavior can be modeled in terms of 'speech flow', 'loudness', and 'intonation', while voice sound characteristics relate to the distribution and intensity of 'overtones' that make up the speakers' individual voice 'timbres'. Speech flow describes the speed at which utterances are produced as well as the number and duration of temporary breaks in speaking. Loudness reflects the amount of energy associated with the articulation of utterances and, when regarded as a time-varying quantity, the speaker's dynamic expressiveness. Intonation is the manner of producing utterances with respect to rise and fall in pitch, and leads to tonal shifts in either direction of the speaker's mean vocal pitch. Overtones are the higher tones which faintly accompany a fundamental tone, thus being responsible for the tonal diversity of sounds. Overtone patterns display large interindividual differences and enable a computerized identification of persons through their voices. On the other hand, stress and affect disturbances modify a subject's overtone pattern in characteristic ways.

Learning to Recognize

Our approach to quantifying speaking behavior and voice sound characteristics for normative purposes relied on 'standardized speech probes' specifically selected for grammatical simplicity: automatic speech, emotionally neutral speech, and emotionally stimulated speech. The 3 types of spoken texts along with repeated assessments at 14-day intervals allowed us (1) to estimate the 'natural' variation of speech parameters over time as a function of language, gender, age, and education, and (2) to determine the sensitivity of speech parameters regarding subtle differences between the emotional states induced by emotionally different speech contents. Emphasis lay on resolving subtle differences between the emotional states induced by different text contents rather than on quantifying the 'emotionality' of the chosen texts through one of the psychological or psycholinguistic models found in the literature [see for example 13–15].

Since speech parameter extraction requires a reproducible subdivision of speech recordings into pauses and utterances ('segmentation'), our segmentation algorithm screened each speech recording for a certain number of intervals without speech signal. These intervals were then used to determine the thresholds for background noise under consideration of a certain 'guard' zone. Subsequently, speech signals were rectified and nonlinearly amplified: low amplitudes were attenuated, and amplitudes above a certain threshold were amplified under the constraint of amplitude ceiling (leading to a cropped waveform).

As language is a critical factor in automatic segmentation, our algorithm was specifically 'trained' through a neural network approach to iteratively optimize the algorithm's free parameters in a language-specific way (multilayer perceptron): (1) we randomly selected 40 recordings (20 males, 20 females) from each language, (2) these recordings were segmented manually and served as reference during the iterative optimization, and (3) the iterative optimization aimed at minimizing the sum of the squared distances between manually and automatically derived segmentation marks under consideration of background noise levels.

Once segmentation was completed, 'spectra' were calculated from discrete Fourier transforms and on the basis of 'pure' utterances with pauses being skipped. We relied on a tonal approach with a quartertone resolution covering 7 octaves in the frequency range of 64–8,192 Hz, so that spectra were comprised of 168 equally spaced quartertones. The tonal approach was chosen because pitch (perceptual quantity) depends logarithmically on frequency (physical quantity). Due to this approach, quartertones were equally spaced on the x-axis, so that rise and fall in pitch could be modeled as linear shifts along this axis.

The longitudinal analyses aimed at assessing stress-induced bodily reactions and affective states as a function of time through a series of 10–20 repeated speech recordings. Focus was laid on quantifying the variation of speech parameters over time ('time course') and on deciding about the significance of the observed variations. Given the distinct individuality of human voices, our approach (1) used each individual as his/her own reference, (2) estimated a 'baseline' by 2nd-degree Gauss approximation, and (3) calculated deviations thereof in both positive and negative direction. Of interest were deviations that exceeded the language-, gender-, and age-specific thresholds given by the normative data. Also, baseline curves were used to compensate for systematic trends due to habituation.

Statistical Analyses

Statistical analyses along with statistical significances were calculated by means of the statistics package SAS 9.3 on a 64-bit Windows system. For the multivariate analyses we used the PROCs STEPDISC, DISCRIM, and GLM combined with subsequent cross-validation.

Results

Language-Specific Segmentation of Speech Signals

The language-specific neural network optimization of the segmentation algorithm yielded a surprisingly consistent set of the algorithm's free parameters. Specifically,

Table 1. Sample composition of the normative study encompassing the languages English, French, German, Italian, and Spanish, and the age groups 18–30, 31–40, 41–50, and 51–65 years

Language	18–30 years	31–40 years	41–50 years	51–65 years	Males	Females	Total
English	52	19	25	21	45	72	117
French	77	19	10	22	44	84	128
German	85	45	33	45	92	116	208
Italian	38	24	27	31	60	60	120
Spanish	39	28	33	24	59	65	124
	291	135	128	143	300	397	697

the sum of squared distances between manually and automatically derived segmentation marks reached distinct minima, even in the case of somewhat 'noisy' recordings ('noisy' recordings with artifactual signals were a minor issue at an overall rate of <5% across all recordings).

Reference Values for Decisions about Deviations from 'Normality'

Using the 5 parameters for the assessment of speaking behavior [pause duration, utterance duration, energy per utterance (loudness), dynamics (variation of loudness), and energy per second], and another 5 parameters for the assessment of voice sound characteristics (mean vocal pitch F0, F0 amplitude, F0 variation, F0 6-dB bandwidth, and F0 contour), detailed analyses revealed distinct, highly characteristic differences between all 5 languages under investigation, irrespective of language family ('stress-timed', 'syllable-timed') (table 1). These distinctive characteristics were remarkably stable over time as demonstrated by correlations between repeated assessments at 14-day intervals (table 2). The underlying data represented paired observations of quantitative traits so that the Pearson product-moment correlation coefficient was used for the assessment of consistency and reproducibility of the quantitative measurements. The correlations were calculated for the subjects' mean values per recording day (mean pause duration, mean energy per syllable, mean vocal pitch, etc.).

Taken together, the chosen speech parameters enabled direct verification of the linguistic prosody theorem stating that 'prosody is a distinctive feature for all languages'. In fact, linear discriminant analyses yielded rates of >85% (counting out loud), >90% (reading out loud an emotionally neutral text), and >90% (reading out loud an emotionally stimulating text) and correctly assigned speakers to their native languages. Analysis of variance carried out separately for the 5 languages under comparison revealed highly significant influences of the speech content on speech parameters ($p < 0.001$), thus underlining the need for standardized speech probes for longitudinal studies. On the other hand, the analysis yielded estimates of the sensitivity of the CSA method as to resolving subtle differences between speech parameters derived from emotionally different speech contents – in particular, when comparing emotionally neutral with emotionally stimulated speech. Results clearly spoke in favor of the emotionally neutral text passage as the 'standard' reading probe for longitudinal analyses[3].

Gender and age explained as much as 15–35% of the observed variance of the parameters assessing the speakers' voice sound characteristics, compared to only 1–3% for speaking behavior parameters. The effect of 'educational level' on speech parameters was generally small, in the range of 1–3%, and did not always reach statistical significance. We used these results to derive language-, gender-, and age-specific thresholds that draw the line been 'natural fluctuations' and 'significant changes' in speaking behavior and voice sound characteristics ('normative data'). Thresholds were determined separately for each language-, gender-, and age-specific stratum and tentatively set to 1.5 standard deviations as such values worked well in our coping behavior study with 2,520 college and university students.

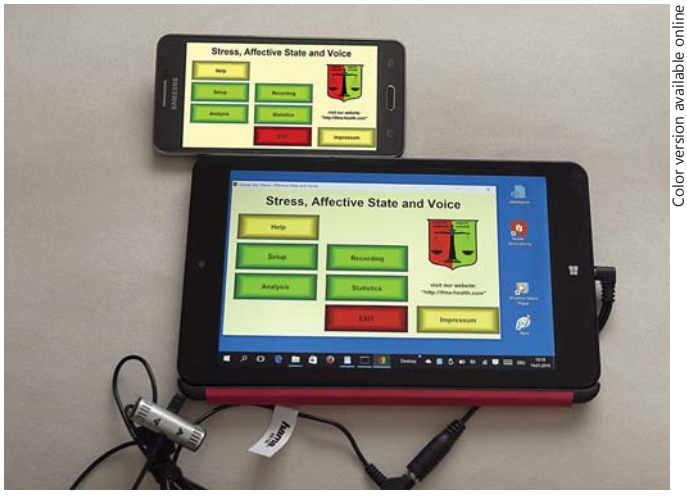
However, these thresholds are of secondary interest for many applications of the 'voice app', in particular when repeated speech recordings are used as 'biofeedback' in attempts to learn to better cope with the ups and downs in daily life. We have learned from our studies with psychiatric patients that a large proportion of patients experienced repeated CSA assessments as a very personal form of therapeutic support. That is, CSA assessments

[3] Prior to starting with our speech studies, we performed a series of extensive pretests with psychiatric patients and healthy subjects stratified according to gender, age, and education. We aimed to find a recording scheme that optimally supported the envisaged analyses while being convenient and acceptable for all types of test persons and patients. The results let us decide in favor of a 'counting-reading-counting' scheme which helped the test person to relax, feel comfortable, and get 'the job done'. Having completed 21 speech studies with more than 6,000 speech recordings from 553 psychiatric patients and 815 healthy controls and repeated assessments under the same experimental setting, we can say that this recording scheme was well accepted by all test persons and patients independent of age, gender, and educational status. Given the large amount of normative data, we see our scheme as kind of a 'gold standard'. The reading probe is available in 5 languages and can be downloaded from the website 'http://www.ifrg.uzh.ch/vox15o.php'.

Table 2. Stability of speech parameters over time in terms of correlations between two recordings at 14-day intervals, computed separately for the languages English, French, German, Italian, and Spanish, and the age groups 18–35 and 36–64 years

Speech parameters	English		French		German		Italian		Spanish	
	18–35 (n = 71)	36–64 (n = 46)	18–35 (n = 88)	36–64 (n = 40)	18–35 (n = 108)	36–64 (n = 96)	18–35 (n = 58)	36–64 (n = 62)	18–35 (n = 57)	36–64 (n = 88)
Pause duration	0.8909	0.8318	0.8663	0.8619	0.8337	0.8849	0.86419	0.88415	0.6651	0.8156
Utterance duration	0.7180	0.6973	0.8576	0.8949	0.8486	0.8945	0.65545	0.85715	0.9468	0.8265
Energy per utterance	0.6030	0.4203	0.5670	0.5966	0.6832	0.5503	0.77054	0.67395	0.1512	0.2478
Dynamics	0.5851	0.3032	0.6215	0.5991	0.6411	0.4719	0.70555	0.65424	0.0876	0.1805
Energy per second	0.6836	0.2924	0.5888	0.5983	0.6715	0.5777	0.76994	0.68778	0.6629	0.5172
Vocal pitch F0 (QT)	0.9369	0.9244	0.9857	0.9639	0.9613	0.9677	0.96187	0.95514	0.7462	0.8959
F0 variation (QT)	0.4900	0.4018	0.2060	0.2065	0.3243	0.4329	0.62396	0.16453	0.4993	0.5652
F0 amplitude	0.7620	0.7289	0.9121	0.8820	0.9002	0.8651	0.79961	0.72495	0.7068	0.8379
F0 6-dB bandwidth	0.5802	0.7164	0.7641	0.7009	0.5570	0.5496	0.63039	0.49903	0.7374	0.7335
F0 contour	0.7662	0.7781	0.8603	0.8536	0.8952	0.8458	0.78909	0.63849	0.7308	0.8048

The higher the correlation coefficient, the lower the impact of environmental factors on this particular aspect of speaking behavior and voice sound characteristics. QT = Quartertones.



Color version available online

Fig. 1. This easy-to-use 'voice app' allows users to monitor affect- and stress-induced behavior over time by means of repeated voice recordings at 1-day intervals. It is currently available for PCs, laptops, tablets, and smartphones under MS Windows (32/64 bit; Windows 7 or higher along with a recent Java Runtime Environment) and Android (4.4 or higher). An iPhone version is in preparation. Available languages are English, French, German, Italian, and Spanish.

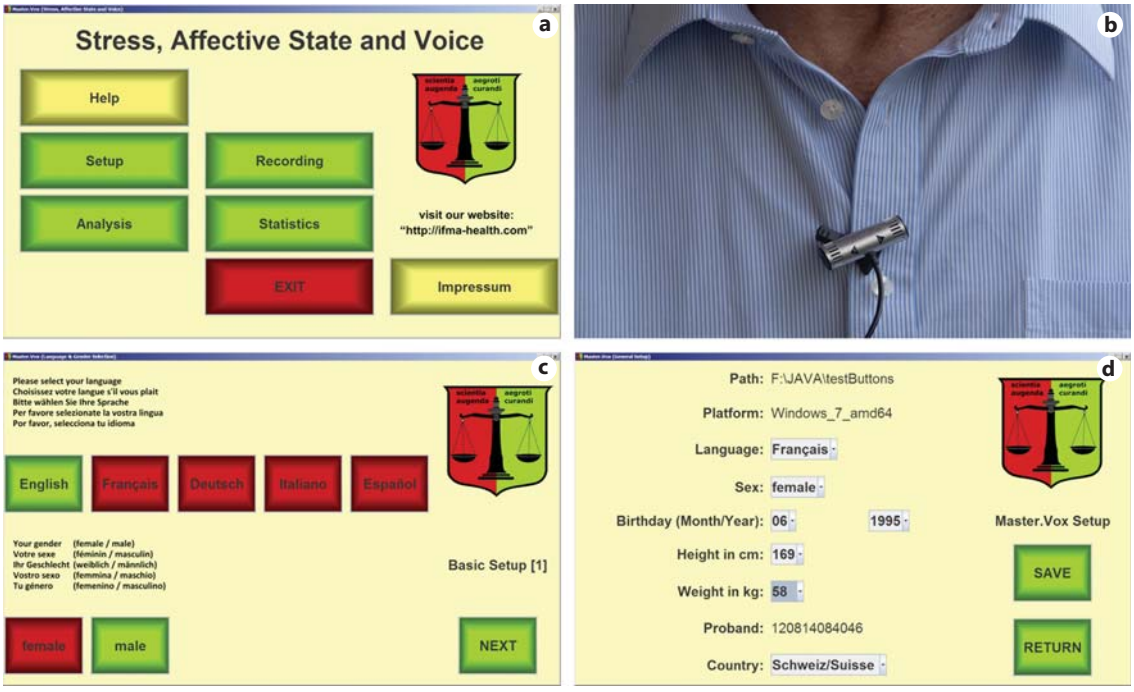
have a therapeutic value by themselves. The same is true for the early detection and prevention of mood disorders: getting involved and doing something about it is the most important step for people under elevated risk of mood disorders.

The 'Voice App'

Using the above-described language-, gender-, and age-specific normative data, we developed a 'voice app' for laptops, tablets, and smartphones that enables self-assessments of stress-induced bodily reactions and affective states through CSA: (1) the recording of speech signals in standardized form, (2) the data management for a series of 10–31 repeated assessments, preferably at 1-day intervals, and (3) the longitudinal analysis of repeated speech recordings which constructs an individual baseline and detects deviations from 'normal' values that exceed the language-, gender-, and age-specific thresholds of 'natural' fluctuations.

The 'voice app' was written entirely in Java and in a largely platform-independent way such that additional speaker languages can simply be 'plugged in' once the respective normative data become available. The 'voice app' is currently available for laptops, tablets, and smartphones under Android (4.4 or higher) or under MS Windows (32/64 bit; Windows 7 or higher with Java Runtime Environment). The iPhone version is in preparation. Available languages are English, French, German, Italian, and Spanish (fig. 1).

Speech signals are recorded at a sampling rate of 48 kHz and a 16-bit resolution either through the internal microphone of the smartphone or tablet (if the quality is acceptable) or through an external microphone attached with a clip to a suitable location of clothing (e.g. HAMA for USD 28; fig. 2b). The same or a similar location of clothing is recommended for repeated assessments. Re-



Color version available online

Fig. 2. The main screen of the 'voice app' displays 4 green buttons (color in online version only): 'Setup', 'Recording', 'Analysis', and 'Statistics' along with 2 yellow buttons 'Help' and 'Impressum' and a red button 'Exit' (**a**). The green buttons provide access to the app's primary sections. When using an external microphone attached with a clip to a suitable location of clothing (**b**), the same or a similar location of clothing should be chosen in repeated assessments. The initial setup procedure lets the users specify 'spoken language' and 'gender' in the first step (**c**) and 'month/year of birthday', 'height', 'weight', and 'country of residence' in the second step (**d**). The correct specification of language, gender, and age is critically important for a reliable quantification of speaking behavior and voice sound characteristics.

cordings are expected to be carried out at a quiet place under comparable experimental conditions. The data are stored as 'WAV' files and are compatible with standard 'WAV' files produced, for example, by the open source program 'Audacity' so that users can easily run their own analyses.

The 'Voice App': Initial Setup

The main screen of the 'voice app' displays 4 green buttons: 'Setup', 'Recording', 'Analysis', and 'Statistics', along with 2 yellow buttons 'Help' and 'Impressum'. The green buttons provide access to the app's primary sections (fig. 2a). During setup, users specify 'spoken language', 'gender', 'month and year of birthday', 'height', 'weight', and 'country of residence' as the quantification of speaking behavior and voice sound characteristics significantly depends on these parameters (fig. 2c, d). The setup is part of the installation procedure, but parameters can be modified at any time.

The 'Voice App': Speech Recordings

The speech recordings comprised 3 pieces of spoken text to assess the speakers' stress-induced behavior and affective state: (1) counting from 1 to 40, (2) reading out loud the standard reading probe, and (3) counting again from 1 to 40. Prior to carrying out speech recordings, the voice app's recording studio must be initialized ('Init' button). At that time, the volume control can be adjusted if necessary (fig. 3a). After volume adjustment, the first recording step (counting) was started by clicking the left 'Start' button and finished by the corresponding 'Stop' button (fig. 3b). This turned the recording LED from red to green and activated the 'Start' button in the middle. The second recording step (reading out loud) could then be started and finished by the corresponding 'Stop' button (fig. 3c). The recording LED turned from red to green and activated the right 'Start' button for the third recording step (2nd counting). Upon completion, users had the possibility to repeat one or several recordings by selecting the respective 'Start' buttons again (fig. 3d). The 'Exit' button leads back to the main screen.

Color version available online

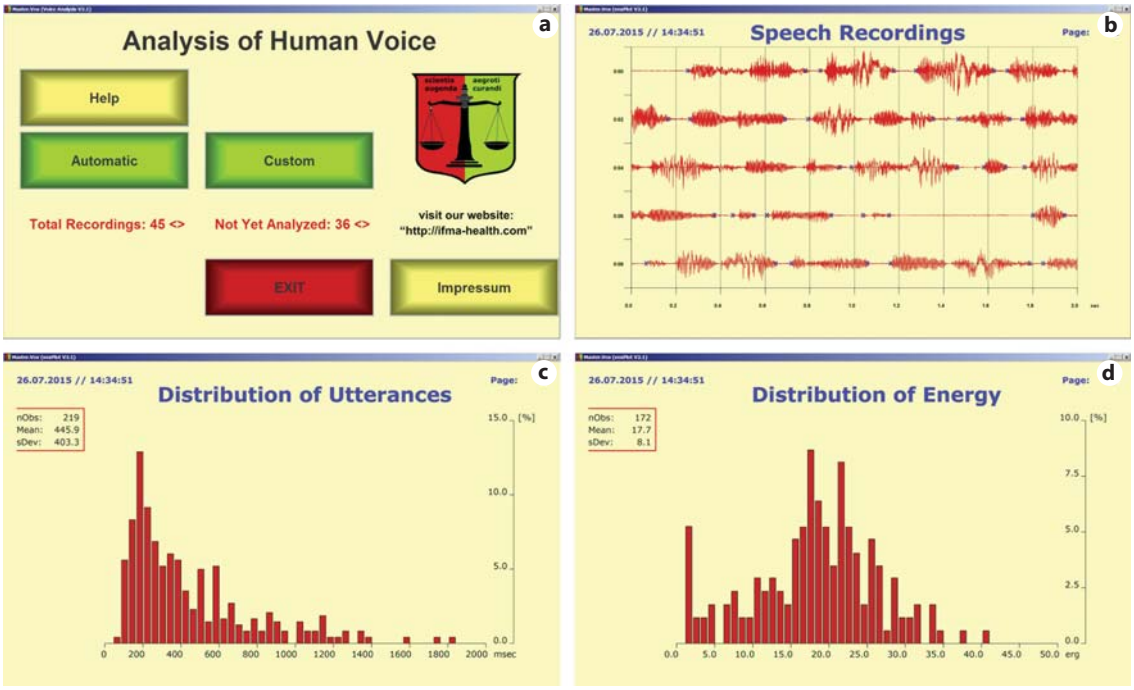
Fig. 3. When the voice app's recording studio is initialized ('Init' button) the 'test' button appears which lets users adjust the volume control by means of some test speech such that the observed maximum level on the volume meter comes close to 95% (**a**). The first recording step (counting from 1 to 40) is started by clicking the 'Start' button on the left and finished by the 'Stop' button (**b**). This turns the recording LED from red to green and activates another 'Start' button in the middle. The second recording step (reading out loud a standard text) can then be started and finished by the 'Stop' button (**c**). The recording LED turns from red to green and activates the 'Start' button for the third recording step (counting again from 1 to 40). Upon completion of the 3 steps, users may repeat each of the recording steps by selecting the respective 'Start' button (**d**).

The 'Voice App': Parameter Extraction (Analysis)

The 'voice app' distinguishes between (1) the analysis of single assessments ('Analysis') which yields a set of speech parameters for each assessment, and (2) the analysis of a set of repeated assessments to evaluate the variation of speech parameters over time ('Statistics'). Single assessment analyses are carried out either automatically ('Automatic') or customized ('Custom'). The customized analysis provides access to details of speaking behavior and voice sound characteristics (fig. 4a). For example, all intermediate steps of the segmentation process can be inspected visually through 'time series plots' (fig. 4b). Such plots give an impression of the data quality, the signal-to-noise ratio, and may reveal the presence of unwanted acoustic artifacts from external sources.

Upon completion of the parameter extraction, the intrinsic properties of speech recordings were visualized by another set of plots ('distribution plots') which enable users to explore the scattering of speech parameters around mean values. In fact, relaxed speakers typically produce a large variety of pauses of different lengths, thus improving understandability while modulating verbal and nonverbal content of speech. This is in contrast to speakers under stress who tend to speak in a more monotone, automatized way. Also, relaxed speakers vary the speed by which they produce utterances to make the presentation more attractive and interesting. People under stress would not do this. Utterance variation plots reveal such details (fig. 4c). Moreover, relaxed speakers typically vary loudness to make a presentation more animated and lively. The width of the variation of loudness (energy) is a direct measure of dynamic expressiveness (fig. 4d).

Spectrum plots revealed the combination and intensity of partial tones that constitute the sound of speech segments (fig. 5a). The first maximum represents the speaker's mean vocal pitch F0, the other maxima show the higher harmonics of F0 (overtones) that make up the 'sound' of a voice. The width of the variation of mean vo-



Color version available online

Fig. 4. Parameter extraction can be carried out either automatically ('Automatic') which is in most cases the preferred method, or in a customized way ('Custom') which yields more detailed insights into speaking behavior and voice sound characteristics (**a**). The segmentation process can be inspected visually through plots from consecutive 20-second epochs. Plot options are 'Time Series' (**b**), 'Low Pass Filter', and 'Marks'. Relaxed speakers typically vary the speed by which they produce utterances to make the presentation more attractive and interesting. People under stress would not do this (**c**). Also, relaxed speakers vary loudness to make a presentation more animated and lively. The width of the variation of loudness (energy) is a direct measure of dynamic expressiveness (**d**).

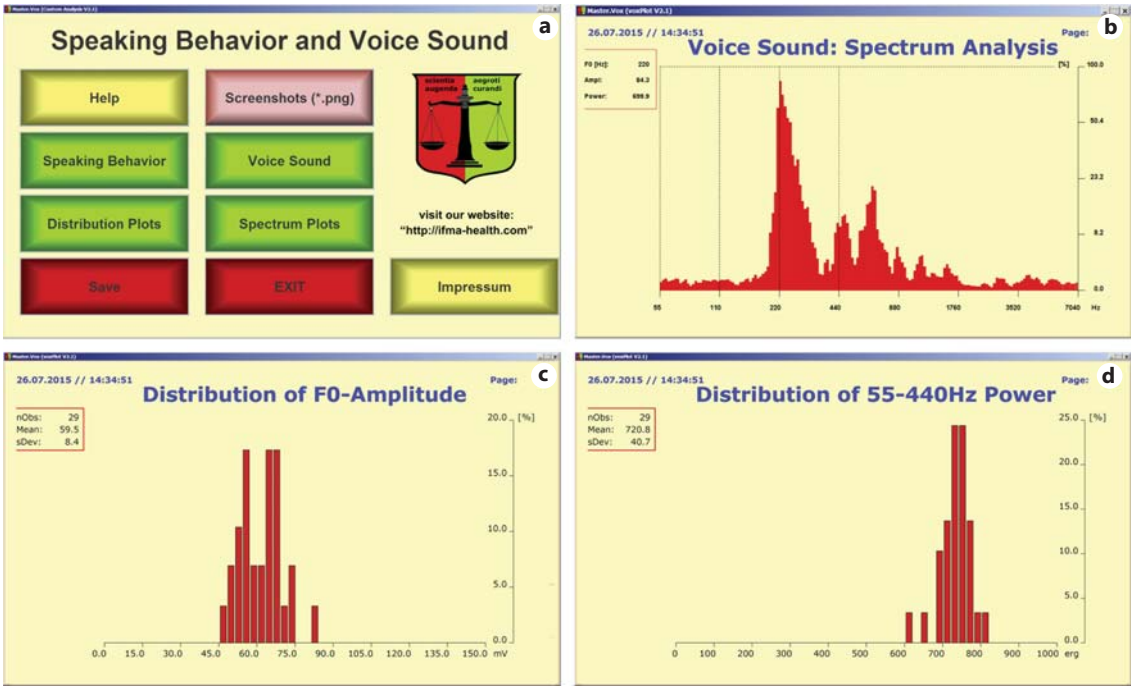
cal pitch is as a direct measure of intonation: the 'broader' the variation the 'richer' the intonation. The spectrum of a female speaker in figure 5b displayed a mean vocal pitch of 220 Hz, which is exactly 1 octave above that of the average male speaker (110 Hz). Higher harmonics were the octave at 440 Hz and the quint at 660 Hz.

The variation of 'F0 amplitude' is an indicator of the 'richness' of a voice sound. A narrow distribution often means deficiency of emotions and empathetic feelings. By contrast, a broad distribution suggests a lively and mindful person (fig. 5c). The variation of '55–440 Hz power' was another measure of the tonal 'richness' of a speaker's voice. Reduced variation indicated a more monotonous presentation of utterances caused, for example, by sorrowfulness, weariness, or fatigue (fig. 5d). In this context it is worth noting that persons under chronic stress tend to speak a halftone above their 'natural' mean vocal pitch and tend to have a 'sharp', sometimes 'metallic' voice sound. Voices regain their 'normal' timbre when stress is reduced and the overall mood brightens.

The 'Voice App': Longitudinal Analyses (Statistics)

For each of the 36 test persons of the self-assessment pilot study individual baselines were calculated along with estimates of the underlying free parameters. Specifically, the algorithm was trained to deal with the various forms of missing data, and to blank out acoustical artifacts that occasionally showed up in the self-assessment recordings. Given the relatively small number of test persons and the interindividual differences in the quality of self-assessments, current optimization results must necessarily be a first approach so that a much larger sample is certainly a must.

Time course plots demonstrated the stability of speech parameters over time along with their 'natural' variation (fig. 6a). These plots provided information about key quantities such as 'pause duration/loudness', 'energy/dynamics', 'vocal pitch/intonation', and 'F0 amplitude/55–440 Hz power'. Despite their remarkable stability over time (fig. 6b), the speech parameters 'pause duration' and 'loudness' occasionally showed a systematic trend toward shorter pauses and greater loudness when speakers got



Color version available online

Fig. 5. Spectrum plots reveal the combination of tones that constitute the sound of speech segments (**a**). This spectrum of a female speaker displays a mean vocal pitch of 220 Hz which is exactly 1 octave above that of the average male speaker (110 Hz). Harmonics are octave at 440 Hz and quint at 660 Hz (**b**). The variation of 'F0 amplitude' is an indicator of the 'richness' of a voice sound. A narrow distribution often means deficiency of emotions and empathetic feelings. By contrast, a broad distribution suggests a lively and mindful person (**c**). The variation of '55–440 Hz Power' is another measure of the tonal 'richness' of a speaker's voice. Reduced variation means a more monotonous presentation of utterances caused, for example, by sorrowfulness, weariness, or fatigue (**d**).

used to the test procedure (fig. 6c). Our analyses automatically compensated for this effect.

In contrast to healthy subjects, patients suffering from mood disorders speak in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering or whispering. During recovery, patients regain their 'normal' speech flow, energy, and dynamic expressiveness (see fig. 7). By contrast, figure 6d shows a test person's time course of 'pause duration' and 'loudness' over a period of 14 days, suggesting that the person's speaking behavior remained virtually unchanged over the entire observation period with the only exception of day 12 where longer pauses and a lower voice indicated sleepiness (no recordings were available on days 4 and 13).

Self-Assessment Pilot Study

Our 14-day self-assessment study with repeated assessments at 1-day intervals was carried out in the home environments of 18 students under exam stress (Zurich: German) and 18 unemployed adults (Valencia: Spanish). We found the CSA method to work surprisingly well in the majority of cases. Stability and 'natural fluctuations' of speech parameters turned out to be comparable to those of the normative study. Incomplete data were due to missing assessments on one or several days where test persons omitted speech recordings for unknown reasons. The missing assessments were noncritical as they were isolated and did not compromise baseline estimation. Most test persons exhibited relatively 'flat' parameter curves over the 14-day period with a broad range of interindividually different baselines. In nearly 20% of test persons we observed some 'habituation effects', for example, a continuous decrease in pause duration combined with a continuous increase in loudness.

Discussion

CSA methods have been successfully used in psychiatry for more than 2 decades to (1) quantify speech dysfunctions among patients suffering from severe mental disorders [4], (2) 'objectively' measure the severity of psy-

Color version available online

Fig. 6. Time course plots reveal the stability of speech parameters over time along with their 'natural' variation (**a**). Despite their remarkable stability over time (**b**), the speech parameters 'pause duration' and 'loudness' often show a systematic trend toward shorter pauses and greater loudness when speakers get used to the test procedure (**c**). **d** A test person's time course of 'pause duration' (red bars; color in online version only) and 'loudness' (green bars) over a period of 14 days (no assessments on days 4 and 13). The figure suggests that the person's speaking behavior remained virtually unchanged over the entire observation period with the only exception of day 12 where longer pauses and a lower voice indicated that the test person may have been sleepy.

chiatric syndrome scores, (3) determine the time point of response to treatment, and (4) monitor the transition from 'affectively disturbed' to 'normal' among patients under treatment [7, 8, 16–23]. Our CSA method has been developed by assessing patients at 2-day intervals by means of standardized psychopathology ratings (observer ratings) and independent 2- to 3-min speech recordings, with raters being 'blind' to speech-recording results and speech-recording technicians 'blind' to psychopathology rating results. For 65–75% of patients suffering from major depressive disorders, we found a close correlation of $r \geq 0.8$ in single-case analyses between HAMD-17 scores and voice sound characteristics (fig. 7). This correlation was much lower and did not always reach statistical significance when patients got 'stuck' in their recovery after initial improvement or exhibited an irregular pattern of HAMD-17 scores over the observation period (typically 25–35% of patients with major depressive disorders).

From a methodological point of view, the CSA method works equally well in self-assessments when monitoring the transition from 'normal' to 'affected' among subjects at risk of affective disorders. However, two problems arise in this context: how to identify those 10% of subjects of the general population who may be at risk, and how to let them carry out standardized speech recordings? As college and university students are a primary target population of our project, we addressed the first problem by developing an online screening tool through a comprehensive study of 2,520 students from 6 colleges and universities in the US, South America, and Europe. This screening tool enables the quantification of basic coping behavior in a cost-efficient and reliable way as well as the 'early' detection of students with insufficient coping skills under chronic stress who may be at risk of physical and mental health problems [1, 2].

The second problem was addressed by translating the patient-oriented, lab-based CSA method into an interactive self-assessment 'voice app' for laptops, tablets, and smartphones. Using a series of 10–31 repeated assessments, the 'voice app' evaluates speaking behavior and voice sound characteristics as a function of time. This is

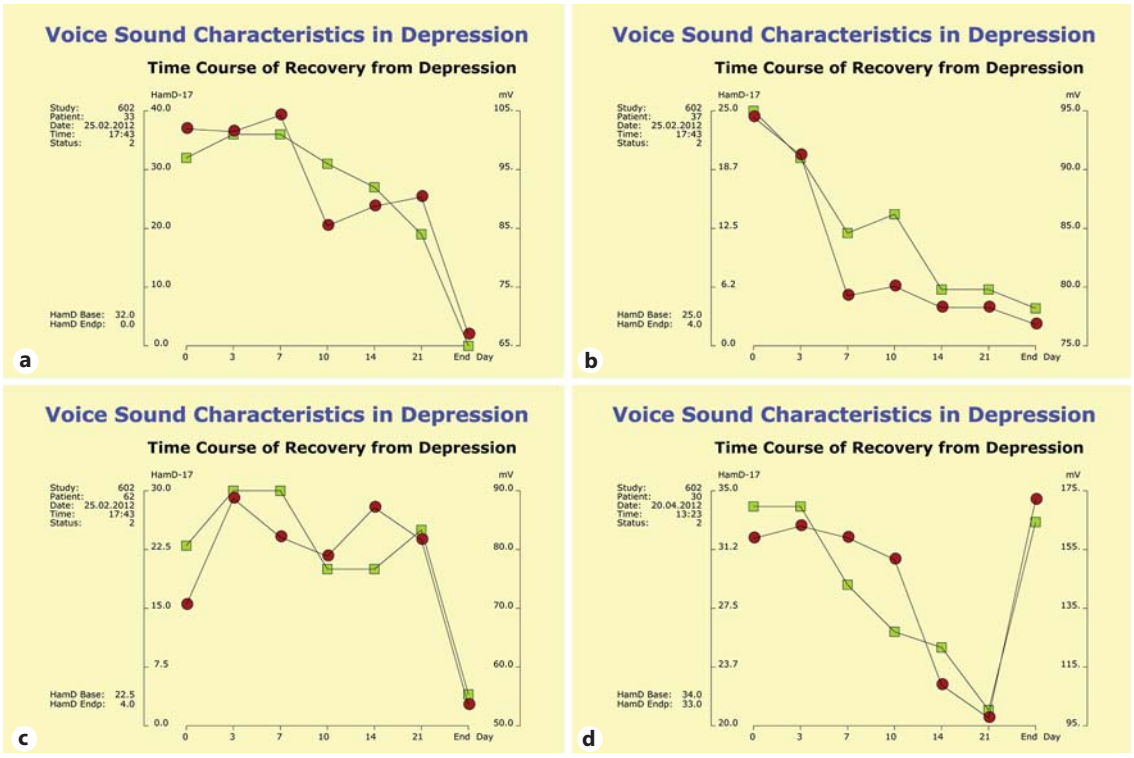


Fig. 7. Time course of recovery from depression under therapy as reflected by 'HAMD-17' scores (green squares; color in online version only) and voice sound characteristics (red circles) from repeated assessments at 2-day intervals along with a final assessment at study end. In clinical studies one typically observes early responders (**a**, **b**), late responders (**c**), and nonresponders (**d**). In 65–75% of cases HAMD-17 scores and voice sound characteristics exhibit a close correlation of $r \geq 0.8$ in single-case analyses. Correlations are much lower and do not always reach statistical significance when patients get 'stuck' in their recovery after initial improvement (some 30% of cases).

particularly helpful for patients who recently recovered from major depressive disorders and may be at risk of relapse. These patients are a target population for the 'voice app', which watches for longer persisting deviations from baseline (>2 days) so that patients can benefit from early interventions prior to the development of clinically relevant symptoms.

College and university students experience significant levels of chronic stress over quite a long time, which can aggravate preexisting psychiatric conditions or trigger the development of new mental health problems or other noncommunicable diseases. Using the 'voice app' means getting involved. And doing something about insufficient coping behavior is the most important step for students at elevated risk of physical and mental health problems.

Specifically, the interactive 'voice app' presents results in such a way that they are directly interpretable by users ('biofeedback'). For example, relaxed speakers typically produce a large variety of pauses of different lengths, thus improving understandability, whereas speakers under stress tend to speak in a more monotone, automatized way. Similarly, relaxed speakers vary the speed by which they produce utterances to make the presentation more attractive and interesting, as they vary loudness to make a presentation more animated and lively. People under stress would not do this. The variation of 'F0 amplitude' is an indicator of the 'richness' of a voice sound. A narrow distribution often means deficiency of emotions and empathetic feelings, whereas a broad distribution suggests a lively and mindful person. Persons under chronic stress tend to speak a halftone above their 'natural' mean vocal pitch and tend to have a 'sharp', sometimes 'metallic' voice sound. Thus, the 'voice app' provides valuable means that allow users to learn to better control stress-related bodily reactions ('biofeedback'). The corresponding thresholds are of secondary interest in this context.

At this stage of the project it is not yet possible to give detailed guidance about the clinical relevance of varia-

tions that persist over more than 2 days because we do not yet have a sufficiently informative database for the target populations. This project is about prevention so that 85–90% of test persons of the general population will display 'natural' variations. In consequence, our current 'voice app' studies start with 400 college/university students in order to identify 40–60 students that may be 'risk cases' with more pronounced and longer persisting deviations from baseline. Moreover, we count on the philanthropy of 'voice app' users when we ask them to upload their voice recordings to our server (in a strictly anonymous way) and to share these data with the research community, thus helping to further improve the 'voice app'.

Key to this project was a normative speech study (5 major languages; n = 697). The sample included 298 males and 399 females covering 4 age groups (18–30 years: n = 291, 31–40 years: n = 135, 41–50 years: n = 128, and 51–65 years: n = 143) while being stratified with respect to education (4 categories: remedial, junior high, high, and college or university). All test persons presented 3 speech types twice at 14-day intervals (automatic speech, emotionally neutral speech, and emotionally stimulated speech). Thus, the study design provided a sound basis for reference data and allowed us (1) to reproducibly quantify the subtle differences between speech parameters related to emotionally different speech contents, (2) to estimate the 'natural' within-subject variation of speaking behavior and voice sound characteristics for each language-, gender-, and age-specific stratum, (3) to estimate the 'natural' between-subject variation of speaking behavior and voice sound characteristics for each language-, gender-, and age-specific stratum, and (4) to derive language-, gender-, and age-specific thresholds that can be used in single-case analyses – where each subject serves as his/her own reference – in order to decide about the significance of changes relative to this reference.

During program development, the app's general user interface (GUI) was extensively tested by potential users so that specific feedback led to an uncomplicated, easy-to-use tool that works equally well for males and females in the age range of 18–65 years (except for persons with reading problems). The results of our self-assessment pilot study underlined the feasibility of the CSA method in home environments, though background noise may occasionally be a limiting factor. A crucial point, however, is the necessity of regular assessments, which requires self-discipline and motivation, particularly in times of negative emotions.

We observed some 'habituation effects' in nearly 20% of the test persons of our longitudinal study with repeated assessments at 1-day intervals. Systematic trends can easily be detected visually by every 'voice app' user. We did not see such effects among patients with repeated assessments at 2-day intervals and do not have a plausible explanation for this. It may indicate lack of interest or that the test person was tired on the study, or had continuously improved reading skills. Though the 'voice app' has implemented a compensation for linear trends, this problem needs further attention. We are working on a solution (for details, see http://ifma-health.com/Left04k.php).

We do *not* use the self-assessment CSA method to classify psychological states or psychiatric disorders nor to derive clinical diagnoses. Attempts to predict 'depression' or 'suicidality' (whatever this might be) can be helpful for populations like U.S. veterans, where the lack of specificity of simple speech analysis approaches like the 'reduced vowel space' method [20] (featuring false-positive and false-negative prediction rates in the range of 20%) may be less relevant, provided the prediction results are intended as information to health care providers only. However, the idea of labeling users with abstract, error-prone, and sometimes frightening diagnoses in self-assessment tools is inappropriate and ethically inacceptable – the more so, as diagnostic systems in psychiatry do not offer clinicians reliable guidelines for therapy and prognosis for a particular patient, so that it is currently impossible to make any prediction of how a particular patient will respond to a particular therapy.

Given the sensitivity of the CSA method regarding the detection of stress-induced behavior and affective states, it is worth noting that short-term, often stress-induced fluctuations in mood are constituents of human life. If, however, significant deviations persist over a longer time period then it might become necessary to do something about it, for example, go for a 20-min walk, ride a bicycle, or do some sport on a regular basis. We have learned from our studies with psychiatric patients [7, 8, 16, 21, 22] that a large proportion of patients experience repeated CSA assessments as a very personal form of therapeutic support. That is, CSA assessments have a therapeutic value by themselves. The same is true for the realm of 'early detection and prevention of mood disorders': Getting involved and doing something about it is the most important step for people under elevated risk of mood disorders. Our low-cost[4], easy-to-use 'voice app' might play a significant role in this context.

[4] Available soon in App Stores for USD 3.99.

Conclusions

We have successfully developed and tested a self-assessment CSA method for the detection of transitions from 'normal' to 'affected' in subjects of the general population in the broader context of mood disorders. Our easy-to-use 'voice app' evaluates sequences of 10–20 repeated assessments and watches for affect- and stress-induced deviations from baseline that exceed the language-, gender-, and age-specific thresholds defined by normative data. Specifically, the 'voice app' provides users with stress-related 'biofeedback' and can help to identify that 10–15% subgroup of the general population that exhibits insufficient coping skills under chronic stress and may benefit from early detection and intervention prior to developing clinically relevant symptoms.

Acknowledgments

This project was funded in part through the 7th EU Framework Programme for Research and Technological Development (grant 248544; OPTIMI: http://www.ifrg.uzh.ch/optimi.php).

References

1 Mohr C, Braun S, Bridler R, Chmetz F, Delfino JP, Kluckner VJ, Lott P, Schrag Y, Seifritz E, Stassen HH: Insufficient coping behavior under chronic stress and vulnerability to psychiatric disorders. Psychopathology 2014;47:235–243.

2 Delfino JP, Barragán E, Botella C, Braun S, Camussi E, Chafrat V, Mohr C, Bridler R, Lott P, Moragrega I, Papagno C, Sanchez S, Soler C, Seifritz E, Stassen HH: Quantifying insufficient coping behavior under chronic stress. A cross-cultural study of 1,303 students from Italy, Spain, and Argentina. Psychopathology 2015;48:230–239.

3 McEwen BS, Gianaros PJ: Central role of the brain in stress and adaptation: links to socioeconomic status, health, and disease. Ann NY Acad Sci 2010;1186:190–222.

4 Stassen HH, Albers M, Püschel J, Scharfetter C, Tewesmeier M, Woggon B: Speaking behavior and voice sound characteristics associated with negative schizophrenia. J Psychiatr Res 1995;29,4:277–296.

5 Juslin PN, Scherer KR: Vocal expression of affect; in Harrigan JA, Rosenthal R, Scherer KR (eds): The New Handbook of Methods in Nonverbal Behavior Research. Oxford, Oxford University Press, 2005, pp 65–135.

6 Mayew WJ, Venkatachalam M: The power of voice: managerial affective states and future firm performance. J Finance 2012;67:1–43.

7 Stassen HH, Kuny S, Hell D: The speech analysis approach to determining onset of improvement under antidepressants. Eur Neuropsychopharmacology 1998;8,4:303–310.

8 Stassen HH, Angst J, Hell D, Scharfetter C, Szegedi A: Is there a common resilience mechanism underlying antidepressant drug response? Evidence from 2,848 patients. J Clin Psychiatry 2007;68:1195–1205.

9 Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB: Age of onset of mental disorders: a review of recent literature. Curr Opin Psychiatry 2007;20:359–364.

10 Braun S, Botella C, Bridler R, Chmetz F, Delfino JP, Herzig D, Kluckner VJ, Mohr C, Moragrega I, Schrag Y, Seifritz E, Soler C, Stassen HH: Affective state and voice: cross-cultural assessment of speaking behavior and voice sound characteristics. A normative multicenter study of 577 + 36 healthy subjects. Psychopathology 2014;47:327–340.

11 Cutler A, Oahan D, van Donselaar W: Prosody in the comprehension of spoken language: a literature review. Lang Speech 1997;40:141–201.

12 Arvaniti A: Rhythm, timing and the timing of rhythm. Phonetica 2009;66:46–63.

13 Ververdis D, Kotropoulos C: Emotional speech recognition: resources, features and methods. Speech Commun 2006;48:1162–1181.

14 Schuller B: On the acoustics of emotion in speech: desperately seeking a standard. J Acoust Soc Am 2010;127:1995.

15 Weninger F, Eyben F, Schuller B, Mortillaro M, Scherer KR: On the acoustics of emotion in audio: what speech, music, and sound have in common. Front Psychol 2013;4:292.

16 Lott PR, Guggenbühl S, Schneeberger A, Pulver AE, Stassen HH: Linguistic analysis of the speech output of schizophrenic, bipolar, and depressive patients. Psychopathology 2002;35:220–227.

17 Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics 2007;20:50–64.

18 Mundt JC, Vogel AP, Feltner DE, Lenderking WR: Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 2012;72:580–587.

19 Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri T: A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015;17:10–49.

20 Scherer S, Lucas G, Gratch J, Rizzo A, Morency LP: Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Trans Affect Comput DOI10.1109/TAFFC.2015.2440264.

21 Kuny S, Stassen HH: Speaking behavior and voice sound characteristics in depressive patients during recovery. J Psychiatr Res 1993;27:289–307.

22 Püschel J, Stassen HH, Bomben G, Scharfetter C, Hell D: Speaking behavior and voice sound characteristics in acute schizophrenia. J Psychiatr Res 1998;32:89–97.

23 Stassen HH, Anghelescu IG, Angst J, Böker H, Lötscher K, Rujescu D, Szegedi A, Scharfetter C: Predicting response to psychopharmacological treatment. Survey of recent results. Pharmacopsychiatry 2011;44:263–272.

24 Kuny S, Stassen HH: The Zurich Health Questionnaire (ZQH): quantifying 'regular exercises', 'consumption behavior', 'impaired physical health', 'psychosomatic disturbances' and 'mental health' in the general population through 63 items. Zurich, Psychiatric University Hospital Zurich, 1988 (available on request).